Early Detection Research Network

NASA JPL Jet Propulsion Laboratory California Institute of Technology

Excellence in Computational Biology and Informatics

.

Daniel Crichton
Dan.Crichton@jpl.nasa.gov
Principal Computer Scientist and Program Manager
Director, Center for Data Science and Technology
Principal Investigator, NCI Early Detection Research Network Informatics Center
NASA Jet Propulsion Laboratory, California Institute of Technology

NATIONAL CANCER INSTITUTE

DARTMOUTH MEDICAL SCHOOL 1797

VANDERBILT UNIVERSITY

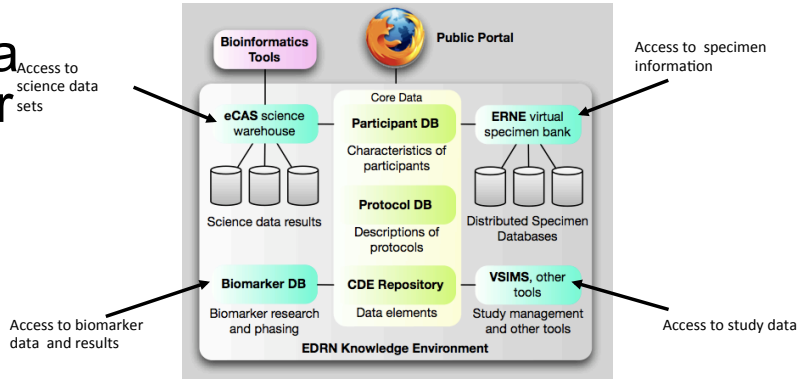
FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

NASA JPL Jet Propulsion Laboratory California Institute of Technology

EDRN Informatics

- NCI/JPL partnered since 2001 to develop a long term distributed knowledge system for the EDRN

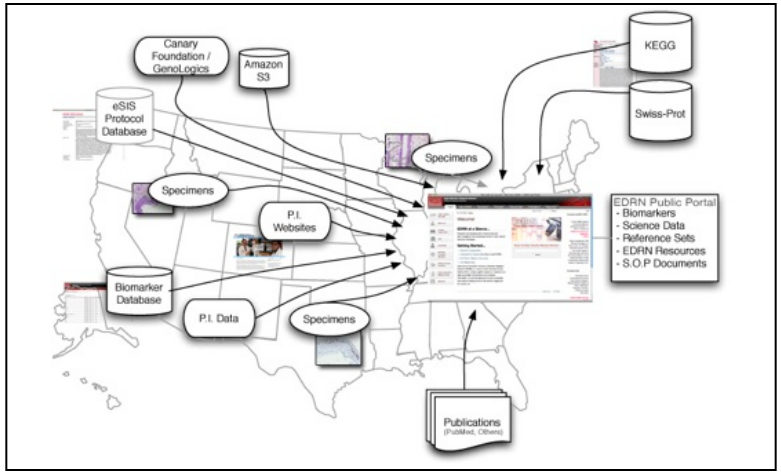
- Significantly leveraged the NASA model
  - Implemented Apache OODT from JPL
  - Architecture and approach
  - Open Source, Data Intensive Science approach
  - 2011 NASA Award for the accomplishment



http://cancer.gov/edrn (operational)
http://edrn.jpl.nasa.gov (beta; emerging capabilities)

Integrated knowledge environment

- Supports capture and access to a diverse collection of distributed sets of information and results
  - Biomarkers
  - Biospecimens
  - Scientific Data Sets
  - Protocols
  - Etc

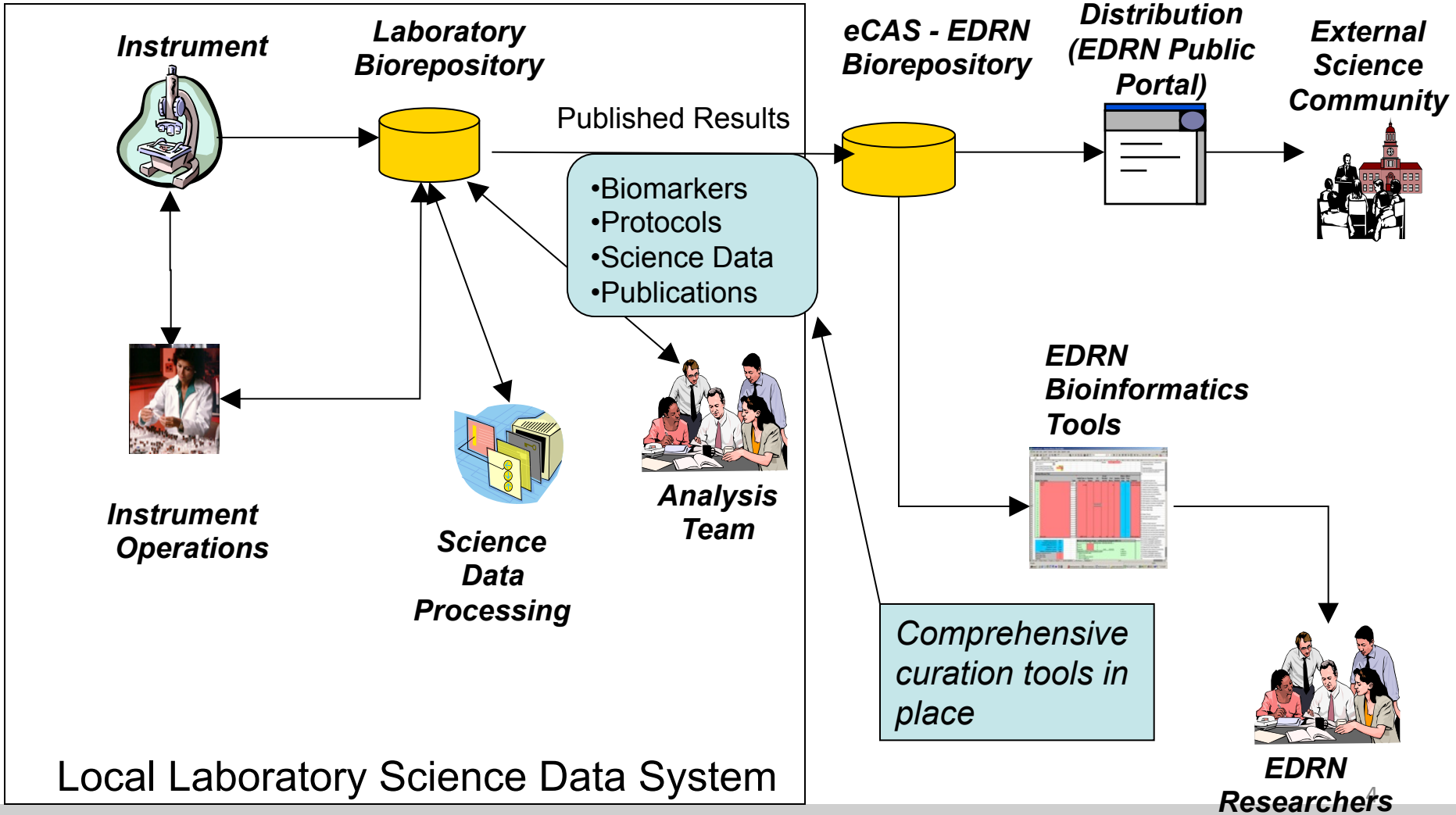


Supporting the Science Data Lifecycle

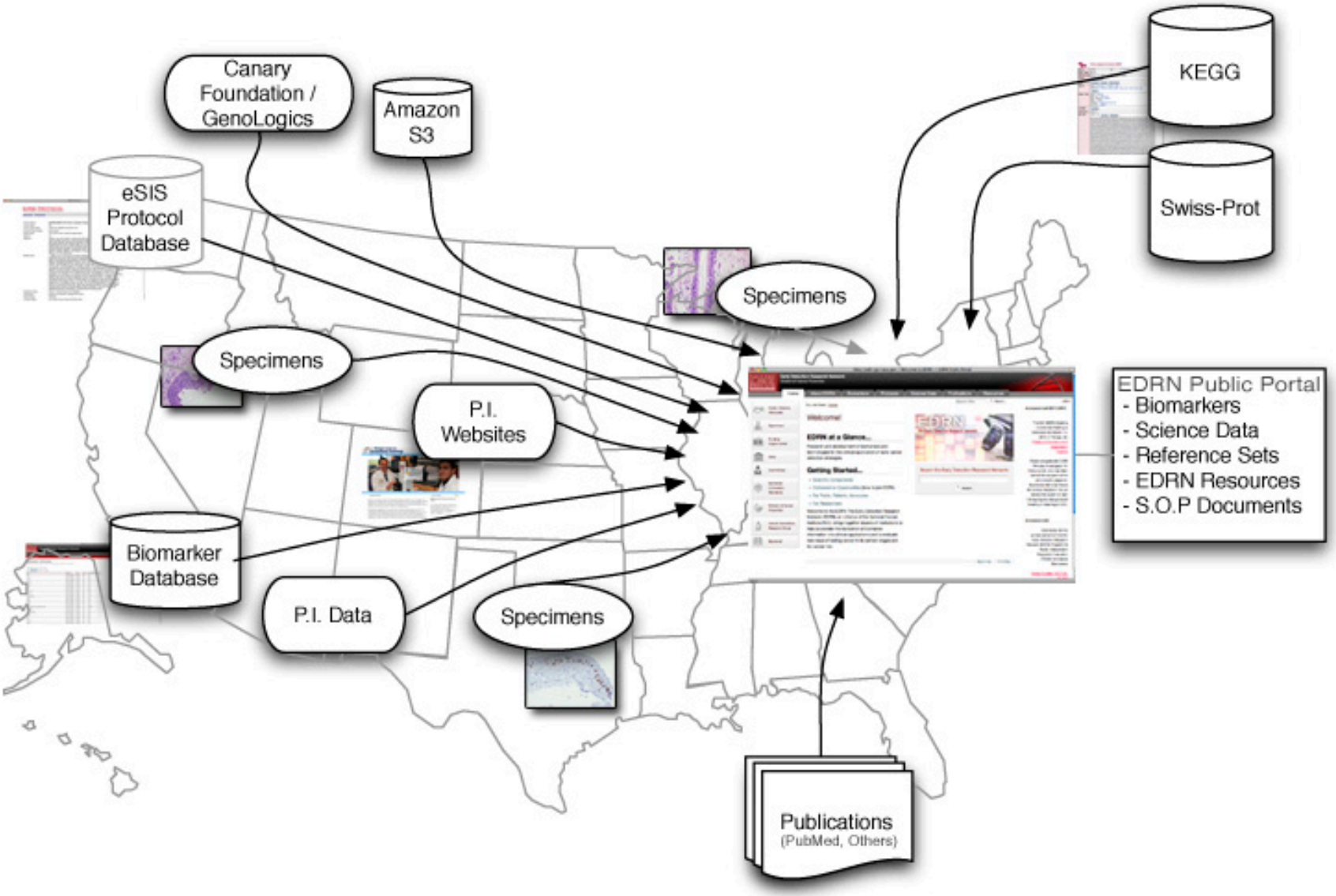
- *Ingestion of data*: Steps for transformation and validation including curation and peer review of the data
- *Cataloging of Structured and Unstructured Data*: Separation of the description (catalog) of data from the physical data storage
- *Data Processing*: Highly validated, scalable pipelines and jobs for remote sensing instruments; versioning of algorithms and data; this can be done by distributed teams prior to submitting to national archives
- *Data Management*: Construction and management of metadata catalogs and data (often distributed); capture of raw and processed data.
- *Data Discovery*: Discovery of data for scientific research
- *Data Access*: Access to the scientific data
- *Data Distribution, Computation and Analysis*: Support for analysis and services (e.g., subsetting) on the data; move towards automated data discovery

Capture of Public Science Data

An Integrated Repository of Public Data Sets

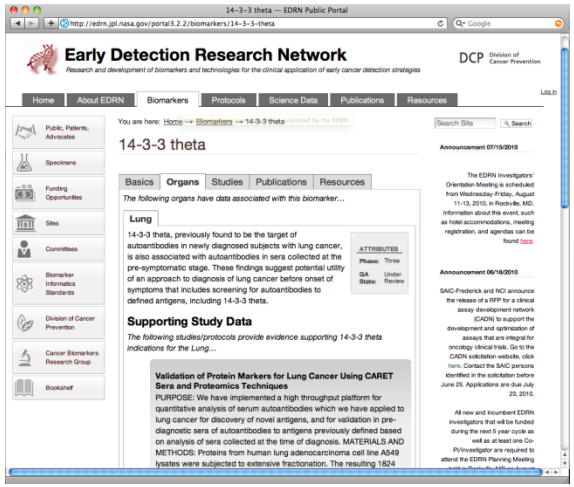


A Virtual, National Integration Biomarkers Knowledge System

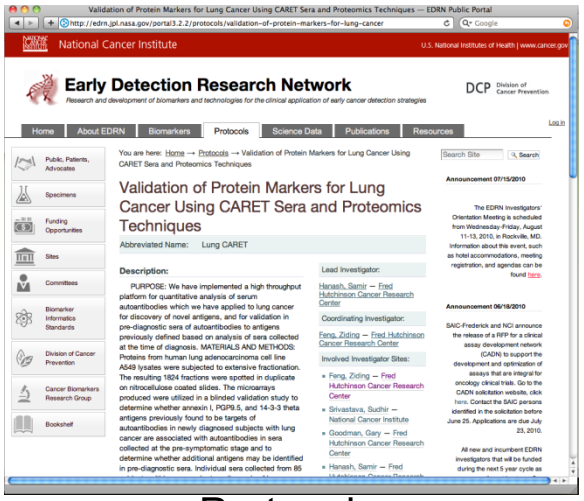


Biomarker Knowledge System: An integrated semantic architecture

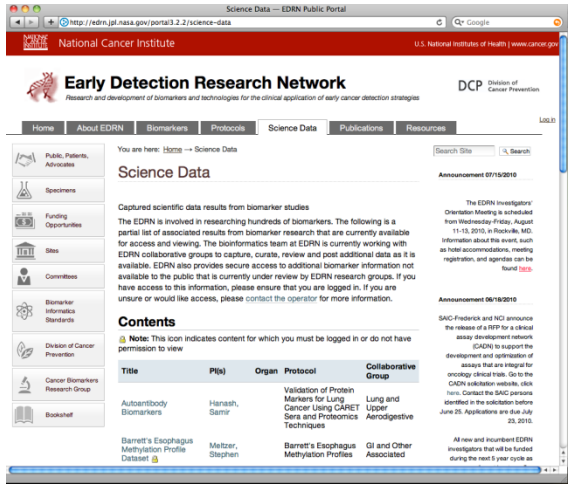
Jet Propulsion Laboratory
California Institute of Technology



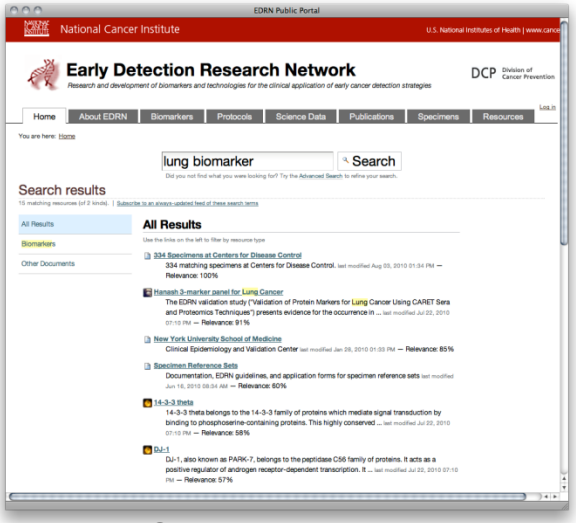
Biomarker Annotations



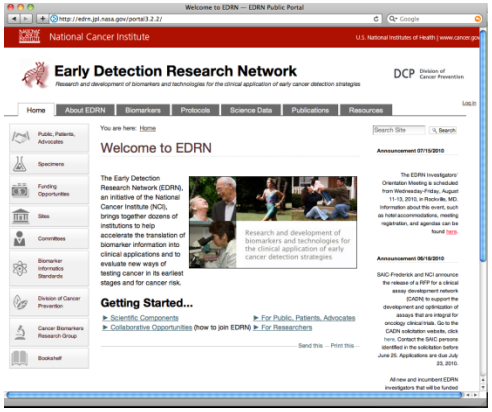
Protocols



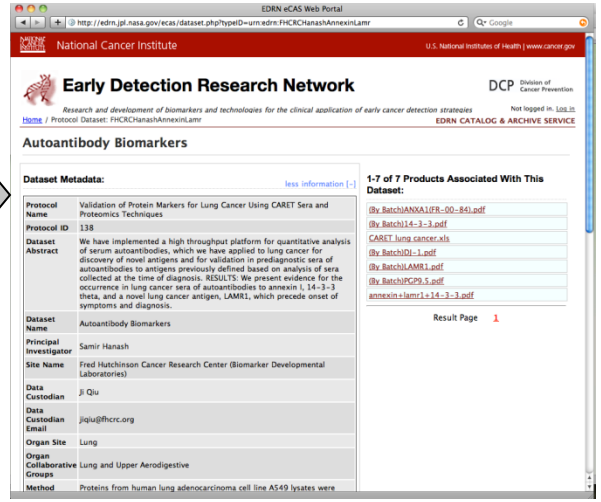
Biomarker Data Results



Specimens



Linked through
Public Portal

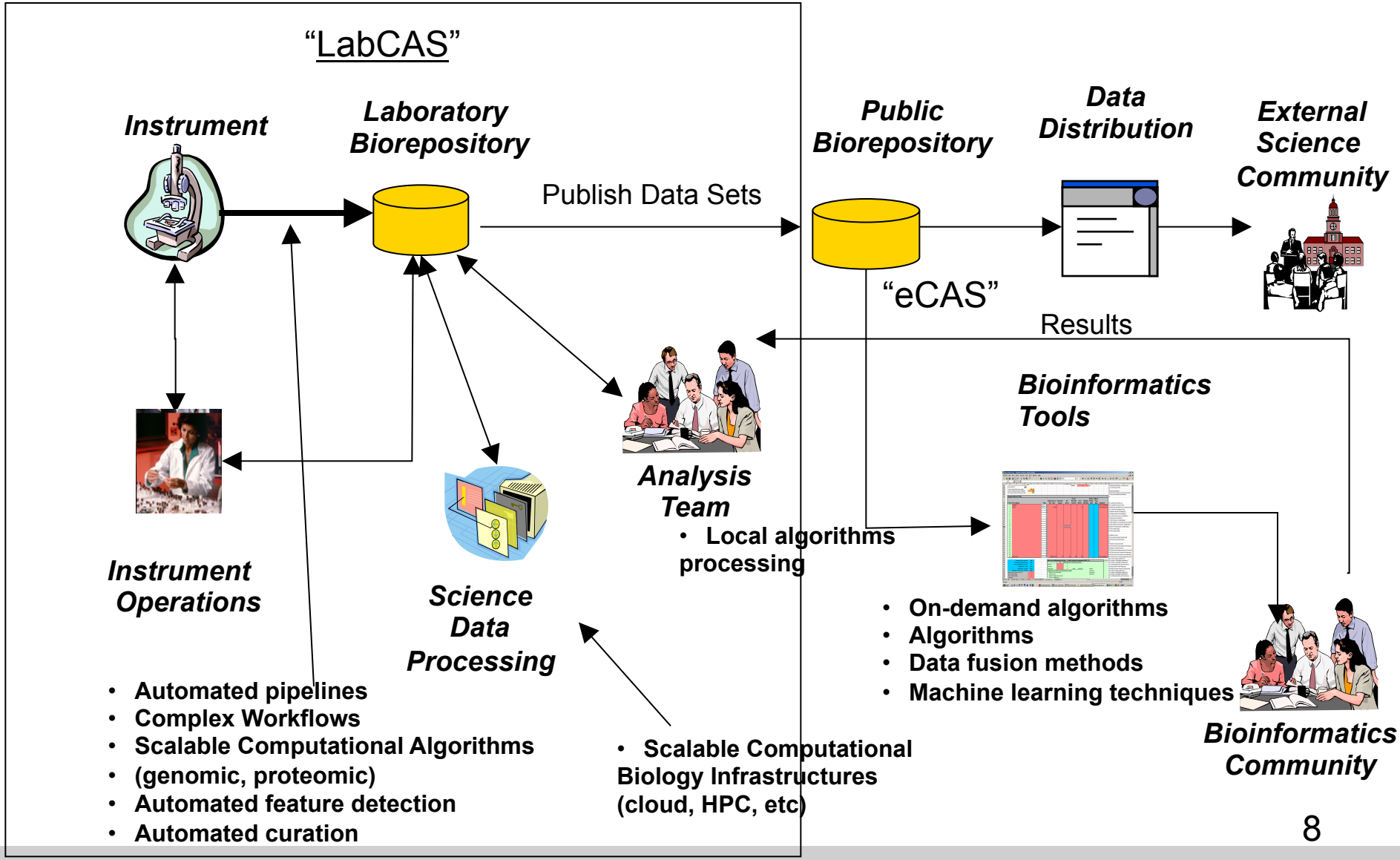


Access to download data

Cancer Biomarker Bioinformatics Workshop

- The EDRN and NASA Jet Propulsion Laboratory held a workshop in May 2013 at Caltech to address informatics and data-driven research in cancer biomarkers
  - http://edrn.nci.nih.gov/cancer-bioinformatics-workshop/cancer-biomarker-bioinformatics-workshop-report-may-2013
  - A major outcome focused on data usability, reproducibility of results, methods and algorithms to systematize data analysis, and scalable computing infrastructures.
- Key Recommendations
  - Systematic approaches to the generation, capture, management of data to enable reproducibility.
  - Increased emphasis on data curation to promote data reuse
  - Automation of data process/analytics software pipelines
  - Data integration and fusion of data from multiple platforms, studies
  - Scalable data infrastructures and repositories
  - Use of big data tools and bioinformatics techniques to scale data analysis
  - Increased training of scientists in the use of computational tools/methods

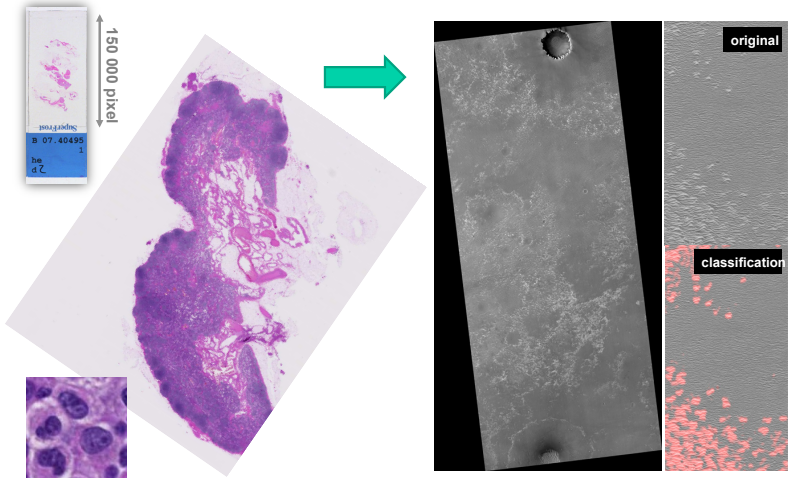
Moving towards data-driven science for cancer biomarkers

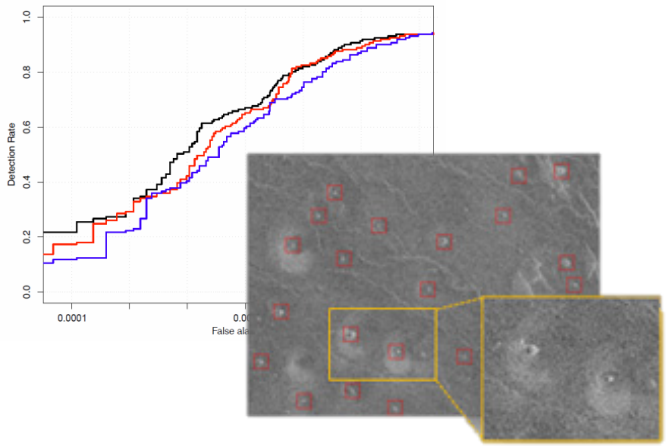


Application of Machine Learning Techniques

Early Detection Research Network

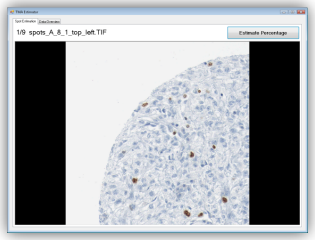
Jet Propulsion Laboratory
California Institute of Technology





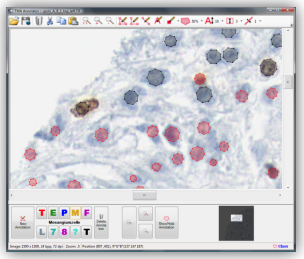
Volcanoes on Venus

TMA Estimator



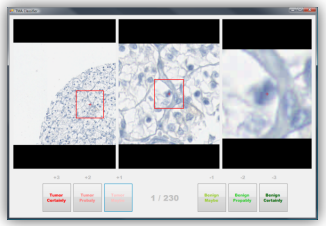
Estimate the Staining on a whole spot

TMA Annotator



Detect nuclei on a whole spot

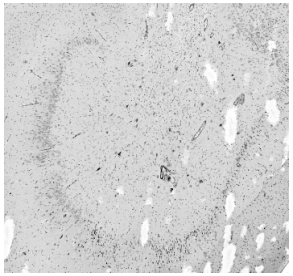
TMA Classifier



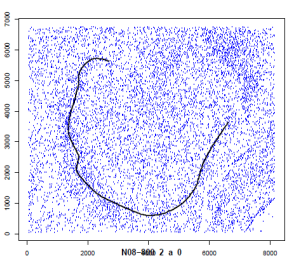
Classify single nuclei into tumor, non-tumor and stained, not-stained

Automated Classification

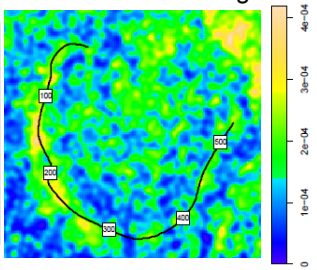
Original Image



Discriminative
Object Detection



Generative
P. Process Fitting



Feature/Object Detection

Early Detection Research Network

Today

Jet Propulsion Laboratory
California Institute of Technology

- Good opportunities to look at collaborations around data-driven computational science approaches
  - Excellent speakers
- Recommend those that are interested to check out the Caltech/JPL Virtual Summer School on Big Data Analytics through Coursera or on the Caltech website
  - Started Sep 2, 2014
  - 1500 people signed up to watch
- I hope you enjoy the session!

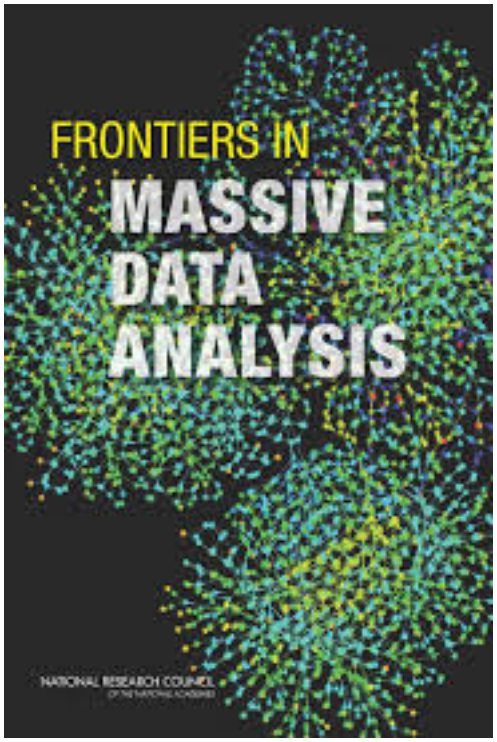
Early Detection Research Network

Backup

Jet Propulsion Laboratory
California Institute of Technology

National Research Council: Frontiers in Massive Data Analsyis

- Chartered in 2010 by the National Research Council
- Chaired by Michael Jordan, Berkeley, AMP Lab (Algorithms, Machines, People)
- Importance of systematizing the analysis of data
- Need for end-to-end approaches to data analysis
- Integration of multiple disciplines
- Application of novel statistical and machine learning approaches for data discovery
- The movement from computation-intensive to data-intensive



Published Sept. 2013